Published in: Grzybek, P., Köhler, R. (eds.): Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday. - New York: Mouton de Gruyter, 2007. p. 449-460.

Two semi-mathematical asides on Menzerath-Altmann's law

Peter Meyer

1 Introduction

Ever since his *Prolegomena to Menzerath's Law* (Altmann 1980) Gabriel Altmann has been much more than a spokesman for the ideas of Paul Menzerath. Finding a mathematical language for the astonishing findings of the German phonetician, he inspired more than one generation's research program on all kinds of quantitative relations between constructs and their components in linguistics and elsewhere, far beyond what Menzerath might have dreamt of (cf. Altmann & Schwibbe 1989). Not surprisingly, then, linguists today usually mean Altmanns's formula $y = ax^b e^{cx}$ or the mathematically equivalent differential equation when talking about "Menzerath's law". In the formula, the independent variable x stands for the length of some linguistic construct such as a word, whereas the dependent variable $y(x)$ signifies the average length of the constituents (say, syllables) of constructs of length x in some linguistic corpus, usually, but not always, a text. The coefficients a, b, c are supposed to be dependent on the corpus at hand; often, as in the present article, the case of $c = 0$ is considered to yield the basic, 'undisturbed' form of the law; see Cramer (2005: 674).

Needless to say, Menzerath-Altmann's (MA) law, if taken as an empirical statement about language, is not simply a mathematical equation. It is, however, difficult to discern precisely which empirical claim the law embraces. For several rather trivial reasons, the deterministic equation in itself cannot possibly be taken to literally express a possible quantitative relationship between constructs and their parts. There is agreement on the fact that the MA law must be interpreted stochastically, although it is not formulated in stochastic terms. Taking the actual practice of working with the law as a starting point, the law should read roughly as follows:

> In linguistic corpora of a suitable type (e.g., linguistically homogeneous texts), a curve of the general form $y = ax^b e^{cx}$ can normally be fitted sufficiently well to the relation between the length of constructs on a certain level (as measured

in number of the constructs' constituents on a hierarchically deeper level of segmentation) and the length of these constituents (as measured in number of constituents on some still deeper level).

Here, the adverb *normally* has to be taken to represent a *normic sentence operator*, that is, it indicates that successful curve fitting is, in a certain sense, the prototypical case, with genuine exceptions still being possible; see Meyer (2006) for more explanation and Lehfeldt & Altmann (2002) for one well-documented exception. A recent proposal on the science-theoretical problems of curve-fitting can be found in Mulaik (2001).

Although there is a lot of literature on the question of how to *explain* the ubiquity of the MA law in language – see Cramer (2005) for a survey –, the law still raises almost as many questions as it is able to answer, as Cramer (2005: 687) rightly puts it. The following remarks intend to present a hypothesis about a class of stochastic conditions under which MA-type relations may arise in any kind of hierarchically segmented linear sequences of some sort of objects (section 2). In section 3, the results of testing some specific variants of this hypothesis with the help of a computer program are reported. Section 4 uses these results to point out that monotonely decreasing MA-type relations between construct and constituent length can hold across an arbitrary number of intermediate hierarchical levels of segmentation, contrary to what has sometimes been assumed. In the final section 5, possible links between the mathematical findings and linguistics are discussed.

2 A mathematical hypothesis about Menzerath-Altmann's law

In order to introduce the basic hypothesis to be dealt with here, some terminology will be introduced now which should, for the time being, be taken by the reader to represent purely formal, computational concepts, although I will illustrate these concepts with linguistic examples. In what follows, I will investigate corpora of linear sequences of elements called *terminal elements*. It is not relevant for our present purpose how many different types of terminal elements there are or whether or not there are any restrictions as to how they may be combined linearly. In linguistics, terminal elements could be the smallest constructional unit under consideration, e.g., phones or phonemes. The maximal linear sequences of terminal elements considered in a certain investigation will be called *level-0 constructs* here; a linguistic example would be words, clauses, sentences or even complete texts, construed

as sequences of, say, phonemes or graphemes as terminal elements. I will assume that every level-0 construct can be segmented into contiguous, non-overlapping sequences of terminal elements, that is, *level-1 constructs*, each of which consists of at least one terminal element (the possibility of empty constituents such as null allomorphs in linguistics is not taken into consideration). Level-1 constructs, in turn, are composed of level-2 constructs, and so on, until a certain level n has been reached whose constructs simply consist of terminal elements that are not grouped together to form constructs of some still lower level. The level of terminal elements itself bears no number. It is admissible for a level-i construct to consist of exactly one level-$(i+1)$ construct. The number $n+1$ of levels is to be considered fixed for a given investigation. The level-j constructs that are part of a given level-i construct will be called its level-j constituents; if $j = i+1$ then the level-j constructs will be said to be immediate constituents of the level-i construct. The *length* of a level-i construct C can be measured in different ways, either as the number of terminal elements belonging to C or as the number of its level-j constituents for any given j, $i < j \leq n$.

Let K be some given corpus of level-0 constructs that are hierarchically structured on $n+1$ levels of analysis. I will say that level i and level j, with $i < j \leq n$, are *MA-related* in K if a curve of the general two-parameter power law form $y = ax^b$ can be fitted sufficiently well to the relation between the length x of level-i constructs ($x \in \mathbb{N}$) and the average length $y(x)$ of level-j constituents of level-i constructs with length x. Unless otherwise stated, I will assume that the length of level-i constructs is measured as the number of its level-j constituents, whereas the length of the level-j constituents is given as the number of its terminal elements.

In the following section, I will try to find evidence for the following hypothesis:

> If, in a sufficiently large and hierarchically structured corpus K of the kind described above, the number of terminal elements of constructs of any level is stochastically independent of the number of the immediate constituents of these constructs, then any two levels in K are MA-related.

The purpose of this hypothesis is to state, in a deliberately informal and imprecise manner, general formal conditions that 'generate' the kind of relations that are described by Menzerath-Altmann's law. No statement is made at this point on the empirical relevance of the hypothesis for, say, linguistics; but see section 5.

3 A computer experiment

A rigorous proof of some sufficiently explicit version of the above hypothesis in the case of multi-level hierarchies is out of the present author's reach. As a kind of substitute, I will present the results of some computer simulation experiments. In the setting of such experiments, precisely formulated versions of the above hypothesis can be tested empirically. As a first step, I will assume that a corpus K of the kind delineated in section 2 is given with the following general properties, to be modified later on:

1. The corpus K consists of a certain number c of level-0 constructs.
2. There are $n+1$ hierarchical levels for the constructs in K.
3. The length of all level-0 constructs in K as measured in number of terminal elements is identically and independently distributed according to a one-displaced binomial distribution with parameters $(l, 2l)$, that is, the mean length of level-0 constructs is l terminal elements, the maximum length is $2l$ terminal elements; for a mathematical commentary on the parameters, see below.
4. The number of immediate constituents of any level-i construct consisting of t terminals ($0 \leq i \leq n$) is distributed according to a one-displaced binomial distribution with parameters (p_i, t): The mean number of immediate constituents of a level-i construct is always a certain fixed number p_i, whereas the actual number of immediate constituents can vary between 1 and t.
5. All admissible segmentations of a construct with some given number of terminal elements into a given number of immediate constituents are equiprobable.

The family of binomial distributions has been selected for no other reason than ease of computation; preliminary results show that the choice of distribution family does not have much influence on the results. In this paper, I specify the one-displaced members of this family through a pair $(mean, maxlen)$ of parameters, where $mean \in \mathbb{R}^+$ is the expected value of the distribution and $maxlen \in \mathbb{N}$ is the maximal possible value of the distribution (the minimal possible value is always taken to be 1). It is easy to see that, for the one-displaced binomial distribution with parameters $(mean, maxlen)$, the probability $P(X = i)$ for the value i, $1 \leq i \leq maxlen$, is given as

$$P(X = i) = \binom{maxlen - 1}{i - 1} \cdot p^{i-1} \cdot q^{maxlen-i}, \text{ with } p := \frac{mean-1}{maxlen-1}, q := 1 - p.$$

The computer experiment consists in (i) producing a hierarchically structured corpus K with the above properties and given test parameters c, l, p_0, p_1, ..., p_n using a suitable programming language with a reliable random number generator; and in (ii) fitting the two-parameter form $y = ax^b$ of the MA law to the data obtained. The experiments I carried out for this paper have been programmed in the high level programming language Python (version 2.3.5), which uses the well-tested Mersenne twister as its pseudo-random number producing algorithm. In Python, the level-0 constructs can easily be represented by the standard mutable sequence data type *list*; for example, the list $[[1,2],[3]]$ represents a construct with two immediate (level-1) constituents. The first of these has two level-2 constructs consisting of one and two terminal elements, respectively, whereas the second one consists of only one level-2 constituent with three terminal elements. The fitting has been done using the nonlinear regression program NLREG by Ph.H. Sherrod (version 5.0). For a numerical evaluation of the regression, I use the determination coefficient R^2, i.e., the proportion of variance explained by the power law function.

In Table 1, sample results of fitting a simple MA-curve ($y = ax^b$) to the relation of level-i construct length and the level-j constituent length dependent on it are shown for different pairs $\langle i, j \rangle$ of levels. The table specifies the regression estimates for the two parameters a and b as given by NLREG and the determination coefficient for these estimates. The results demonstrate that the general hypothesis is borne out with respect to the experimental settings chosen. Similar results are obtained with a wide variety of other choices of test parameters, particularly even with a significantly smaller number c of level-0 constituents. However, no systematic exploration of possible ranges of the parameters $c, l, p_0, p_1, \ldots, p_n$ has taken place so far. Notice that the results of the computer experiment are not weighted as to the relative frequency of construct lengths – for more discussion of this point, see Gieseking (2002). This can cause a very low proportion of outliers, that is, constructs with an unusually high number of immediate constituents, to spoil the regression results. In the results table, where the determination coefficient has been marked with a star (*), very high construct lengths (up to a maximum of 1 percent of the constructs) have been removed from the statistics in order to eliminate the influence of outliers.

In the table caption, the 'number of necessary corrections' refers to a problem that may occur in the process of constructing the corpus. Sometimes, the number t of terminal elements of a given level-i construct may turn out to be

lower than the mean number p_i of immediate constituents. In this case, no one-displaced binomial distribution with parameters (t, p_i) is available; my completely ad-hoc solution consists in simply segmenting such constructs into t constituents consisting of one terminal element each. The number of constructs where such a correction was necessary is annotated in the table. In order to reduce the number of such cases, one should see to it that the mean number of terminal elements of level-n constructs is sufficiently high. In linguistic terms, this could mean to choose terminal elements that represent a very small time interval instead of, say, phones.

Table 1: Results of a sample experiment with parameters $c = 3000$, $l = 700$, $n = 3$, $p_0 = 2.1$, $p_1 = 3.3$, $p_2 = 2.0$, $p_3 = 2.5$. Number of necessary corrections: 7703

MA-relation Between	Estimate for a	Estimate for b	R^2
level 0 and level 1	700.11	−1.0	1.0
level 0 and level 2	696.74	−1.0	1.0
level 0 and level 3	697.46	−1.0	1.0
level 0 and level 4	703.33	−1.0	1.0
level 1 and level 2	325.45	−0.98	1.0
level 1 and level 3	314.05	−0.96	1.0
level 1 and level 4	297.14	−0.94	1.0
level 2 and level 3	98.71	−0.93	1.0*
level 2 and level 4	79.35	−0.79	0.98
level 3 and level 4	40.86	−0.65	0.99*

So far we have looked at one specific, but rather straightforward scenario compatible with the general hypothesis. We will now look at the consequences of two modifications of our experimental setting, starting with condition 4. above. The mean number of immediate constituents of a given level-i construct will now be considered not to be a fixed number p_i any more, but to vary according to the actual number of terminal elements of the construct. The expected value v_i for the length of a level-i construct is $\frac{l}{p_0 \cdot p_1 \cdot \cdots \cdot p_{i-1}}$ terminals. If the actual length is a terminal elements, then this actual length differs from the expected mean length by a factor of $\frac{a}{v_i}$. To vary our experimental setting a little bit and in order to render it perhaps a bit more realistic, we will now assume that the mean number of immediate constituents of a level-i construct with a terminal elements is $p_i \cdot \left(\frac{a}{v_i}\right)^f$. In other words, if the actual length a is exactly equal to the mean length v_i, then the mean number

of immediate constituents continues to be a certain number p_i which is fixed for the given level i. If a is larger or smaller than v_i by a factor $\frac{a}{v_i}$, then the mean number of immediate constituents is also accordingly larger or smaller; the exponent f regulates the 'influence' of the factor $\frac{a}{v_i}$ on the mean number of immediate constituents. Typically, f will be in the range between 0 and 1. If $f = 0$, then there is no influence, as in our first model; if $f = 1$, then the mean number of immediate constituents varies by the same factor as the actual number of terminal elements of the construct in question.

A further change may be proposed as to condition 5. above. If all admissible segmentations of a sequence of t terminals into a given number x of immediate constituents are equiprobable, then the random algorithm de facto favors shorter constituents over longer ones, yielding a clean, monotonely decreasing power law distribution of the number of constructs with respect to the number of terminal elements. This can be changed in an admittedly artificial fashion by producing a fixed number s of equiprobable admissible construct segmentations and choosing among them the segmentation with the lowest variance. By means of this formal maneuver, the number of necessary corrections can be reduced drastically.

The results of a sample experiment with the new test parameters f, s are shown in Table 2. Perhaps surprisingly, the results turn out not to change much even in cases where f is close to 1; in this case at least, the number of immediate constituents is not, strictly speaking, stochastically independent of the construct length as measured in number of terminal elements. It seems to be the case that what is most relevant is that the length of a level-i construct (in number of terminal elements) is "chosen" before its number of constituents is determined.

4 'Indirect' Menzerathian relations

The results of the previous section show that under the experimental conditions chosen *all* hierarchical levels are MA-related with one another with a *negative* value of the estimate for the exponent b, which means an average constituent length monotonely decreasing with increasing construct length no matter whether the hierachical levels of constituents and constructs are neighboring or not. At first sight, this seems to contradict intuitions: When a construct on level i gets longer, its constituents on level $i+1$ become shorter; when the constituents on level $i+1$ become shorter, *their* constituents on

Table 2: Results of a Sample Experiment With Parameters $c = 1000$, $l = 1200$, $n = 3$, $p_0 = 3.1$, $p_1 = 2.3$, $p_2 = 1.9$, $p_3 = 3.6$, $f = 0.6$, $s = 2$. Number of necessary corrections: 0.

MA-relation Between	Estimate for a	Estimate for b	R^2
level 0 and level 1	1199.99	−1.0	1.0
level 0 and level 2	1195.04	−1.0	1.0
level 0 and level 3	1188.02	−1.0	1.0
level 0 and level 4	1139.60	−0.99	1.0
level 1 and level 2	216.16	−0.25	0.86*
level 1 and level 3	115.49	−0.13	0.85*
level 1 and level 4	33.56	−0.07	0.77*
level 2 and level 3	108.50	−0.20	0.89*
level 2 and level 4	34.71	−0.12	0.85*
level 3 and level 4	34.06	−0.16	0.93*

level $i+2$ must, in turn, get longer; therefore, it seems, *longer* constructs on level i condition *longer* constituents on level $i+2$, resulting in a monotonely increasing MA-curve with positive exponent b. This assumption has been put forward several times in the literature and has been backed up by the following mathematical reasoning (Altmann 1983): Let x, y, z be the lengths of constituents on levels i, $i+1$, $i+2$, respectively; then from $y = ax^b$ and $z = cy^d$ we have $z = c\left(ax^b\right)^d = \left(ca^d\right)x^{bd}$; therefore, if $b < 0$ and $d < 0$, the MA-relation between level-i constructs and level-$i+2$ constituents has a positive exponent bd. The above experimental results show that there must be something wrong here. The reason for this is a simple equivocation in the naming of the variables: In the equation $z = cy^d$, the variable y is independent (with values being natural numbers) and signifies the length of level-$i+1$ constructs as measured in number of immediate level-$i+2$ constituents. In the equation $y = ax^b$, in contrast, the variable y (i.e., $y(x)$) is dependent and stands for the *average* length of level-$i+1$ constituents belonging to level-i constructs of a certain length x; here, y is measured in number of *terminal elements*. Therefore, the latter equation cannot be inserted into the first one.

5 Empirical consequences?

So far, the data obtained are the results of a formal game with random numbers. The general hypothesis they support has no other purpose than that of

specifying a type of stochastic *mechanism* that (among other types of mechanisms) generates MA-law-like power law relations in ensembles of hierarchically structured entities of whatever kind. Studying such a mechanism gains empirical interest as soon as real-world phenomena are found to conform at least approximately to the formal description of the mechanism. If an empirical phenomenon can be shown to fulfill the preconditions for the mechanism *and* if independent reasons can be given for this fulfillment, then formal results as the ones presented above are part of a genuine *explanation* of the MA-relation.

Taking the example of Quantitative Linguistics, two different tasks must be tackled. First, one must investigate suitable corpora of linguistic data in order to find out whether or not the mathematical preconditions mentioned above are met to a sufficient degree. For instance, it is an empirical matter whether the distribution of length of *words* in texts of a certain genre in some language can be successfully described as stochastically independent of the distribution of *morpheme* length. I will have to leave this question open here; answering it will require large-scale empirical studies.

Second, if the preconditions are indeed found to be met, then it is necessary to find out why. This enterprise requires a *theoretical* reflection on the empirical concepts involved. In the case of hierachies of linguistic entities, this would mean to look at the principles and premises of segmenting linguistic utterances into constructs of different levels of description. Up to now, such considerations have been mostly absent in quantitative linguistic studies. It should be obvious, however, that we can get no deeper and more principled understanding of empirical regularities such as the MA law unless our explanatory account considers the theoretical foundations of the concepts involved. In the example just alluded to, for instance, the criteria used (mostly implicitly) to segment utterances into (phonological or grammatical) word forms are quite different from those employed to find morpheme or syllable boundaries, a point perhaps most obvious in the American structuralist literature on these topics (Bloomfield 1933). Another way of coming to grips with this issue is to look at some empirical and theoretical aspects of language change and grammaticalization. Let me just hint at two illustrative examples.

Pervasive sound shifts that may radically alter the length and syllabic structure of words are mostly insensitive to the morphemic structure of the word. Thus, in the prehistory of the Goidelic branch of the Celtic languages (Irish, Scottish Gaelic, Manx) almost every other syllable has been lost due to vowel elimination (Thurneysen 1946); a similar process has occurred in

the history of French. Such processes do not affect the number of morphemes of words, at least not immediately; they only affect the length of words and morphemes as measured in phoneme number. In such a case, it is somewhat plausible to assume that the distribution of word length as measured in morpheme number develops independently of the distribution of word length as measured in number of phonemes. A similar conceptual and diachronic independence becomes obvious in the realm of syntax if one investigates the segmentation of phrasal constructs into word forms. Frameworks as different as Langacker's Cognitive Grammar (Langacker 1987), Feilke's investigations on the idiomaticity of syntax (Feilke 1996), and the recently proposed generative framework of Jackendoff & Culicover (2005) have emphasized the conceptual autonomy of phrasal syntagms or constructions vis-à-vis their constituents on the level of individual word forms. Said in a somewhat oversimplified way, the syntax of a language is treated by such accounts as an ensemble of a large number of often at best semi-productive patterns. It is natural to assume, from such a point of view, that the number of constituents a phrasal construct may reasonably be divided into can decrease in the course of time as the internal structure of the phrase gets more and more opaque to the language users. This decrease in number of constituents is, again, in both conceptual and diachronic respects, possibly rather independent of the development of the phonetic or phonological development of such syntagms. Again, what we get is a theoretical possibility to understand why the number of constituents in a phrasal syntagm and the length of these constituents themselves might be MA-related in many languages.

To sum up, I have proposed yet another possible pathway to a *part* of an explanation of the ubiquity of MA-relations and language and elsewhere. It might have the advantage of being independent of (but, of course, not incompatible with) psycholinguistic or cognitive considerations that are not applicable to MA-relations found outside language. Additionally, it is more easily amenable to an explanation of the fact that MA-relations often enough hold even in rather short texts, a fact that is difficult to reconcile with more psycholinguistically oriented explanations. Of course, it is both mathematically and empirically obvious that the mechanism sketched is just one of several co-occurring factors that lead to MA-curves in language; for example, an exponent close to -1 as often found in our computer experiments is not typical of MA-relations in human language texts.

Suffice it to remark here that it is easy enough to modify the experiment in order to obtain corpora where all construct levels are MA-related with

negative power law exponents arbitrarily close to 0. One of several possibilities is to start off with some hierarchically structured corpus of the sort discussed here and then to allow randomly distributed numbers of level-0 constructs to coalesce their immediate constituents into new, longer level-0 constructs. This way, another diachronically and conceptually plausible mechanism comes into play, viz. that of preexisting elements combining to form larger constructs; the mechanism we have looked at before instead takes preexisting elements that get segmented into constituents. It will be the task of a separate publication to demonstrate in more detail how these two basal mechanism can interact in different ways to produce many-level MA-relations.

At any rate, all mathematical musings about Menzerath-Altmann's law are useless for Quantitative Linguistics unless they are filled with linguistic content. This is what we all have been taught by Gabriel Altmann, who, despite having mathematical absolute pitch, continues to be, in the first place, one of the foremost linguists of our times.

References

Altmann, Gabriel

1980 "Prolegomena to Menzerath's law". In: *Glottometrika 2.* Bochum: Brockmeyer, 1–10.

1983 "H. Arens' «Verborgene Ordnung» und das Menzerathsche Gesetz". In: Faust, Manfred; Harweg, Roland; Lehfeldt, Werner; Wienold, Götz (Hg.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik.* Tübingen: Narr, 31–39.

Altmann, Gabriel; Schwibbe, Michael H.

1989 *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Bloomfield, Leonard

1933 *Language.* New York: Holt.

Cramer, Irene M.

2005 "Das Menzerathsche Gesetz". In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Raimund G. (Eds.), *Quantitative Linguistik. Ein internationales Handbuch.* Berlin / New York: de Gruyter, 659–688.

Feilke, Helmuth

1996 *Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik.* Frankfurt/M.: Suhrkamp.

Gieseking, Kathrin

2002 "Untersuchungen zur Synergetik der englischen Lexik". In: Köhler,

Reinhard (Hg.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik.* [http://ubt.opus.hbz-nrw.de/volltexte/2004/279,387–433]

Jackendoff, Ray; Culicover, Peter
2005 *Simpler Syntax.* Oxford, New York: Oxford University Press.

Langacker, Ronald W.
1987 *Foundations of Cognitive Grammar. Vol. I: Theoretical Prerequisites.* Stanford: Stanford University Press.

Lehfeldt, Werner; Altmann, Gabriel
2002 "Der altrussische Jerwandel". In: *Glottometrics*, 2; 34–44.

Meyer, Peter
2006 "Normic Laws in Quantitative Linguistics". In: Grzybek, Peter (Ed.), *Contributions To the Science of Language. Structures of Frequencies and Relations.* [In print]

Mulaik, Stanley A.
2001 "The Curve-Fitting Problem: An Objectivist View". In: *Philosophy of Science*, 68; 218–241.

Thurneysen, Rudolf
1946 *A Grammar of Old Irish.* Dublin: Dublin Institute for Advanced Studies.